THE LONG HOLD: STORING DATA AT THE NATIONAL ARCHIVES

Kenneth Thibodeau, Ph.D.
Director, Center for Electronic Records
National Archives and Records Administration

The National Archives is, in many respects, in a unique position. For example, I find people from other organizations describing an archival medium as one which will last for three to five years. At the National Archives, we deal with centuries, not years. From our perspective, there is no archival medium for data storage, and we do not expect there ever will be one. Predicting the long-term future of information technology, beyond a mere five or ten years, approaches the occult arts. But one prediction is probably safe. It is that the technology will continue to change, at least until analysts start talking about the post-information age. If we did have a medium which lasted a hundred years or longer, we probably would not have a device capable of reading it.

The issue of obsolescence, as opposed to media stability, is more complex and more costly. It is especially complex at the National Archives because of two other aspects of our peculiar position. The first aspect is that we deal with incoherent data. The second is that we are charged with satisfying unknown and unknowable requirements.

The data is incoherent because it comes from a wide range of independent sources; it covers unrelated subjects; and it is organized and encoded in ways that not only do we not control but often we do not know until we receive the data.

The sources are potentially any operation of the Federal Government, or its contractors. The National Archives has been in the business of collecting digital data for two decades. The way we get it is through our authority over all Federal records. Under the Federal Records Act, no agency of the Federal Government can destroy or alienate any Federal record without authorization from the Archivist of the United States, who is the head of the National Archives and Records Administration. Simplistically, the way it works is that agencies tell us what records they have, and we tell them which ones they can destroy when they no longer need them, and which ones must be preserved for posterity. (The definition of Federal record in the law explicitly includes machine-readable files.)

Since 1972, we have reached agreements with agencies that provide for them to transfer to us, and for us to preserve, data from 600 data collections. 573 of them are still active. From these agreements, we have received over 10,000 data files. The rate of transfer has increased dramatically in the last two years: In fiscal year 1988, the National Archives received 167 data files. In FY 1989, 645 files came in, and in FY 1990 729. We anticipate a total of 1400 this year. And in each of the next two fiscal years we expect to receive at least 3000 data files. So we are currently operating at eight times the volume of new files we had three years ago, and we expect at least to double that next year.

Those numbers are very encouraging, but the overall picture is rather bleak. If we look at all of the data which was scheduled to arrive in the last twenty years, from those 600 data collections, we have received less than 7% of the transfers which should have been made. We have recently completed development of a system to generate dunning letters to agencies who fail to transfer data as scheduled, and to track each case to completion. But this system creates additional problems. If I implement it as planned, on a governmentwide basis, we would need to increase

our capability to handle new files, not by doubling current capacity, but by increasing it more than six times. And to handle the backlog of data which should have come in before now, I would need at least 10 times our current capacity.

The past gives us pause. But the future is a brave new world. At least it requires a degree of bravura just to glance in that direction. We have underway a study which is looking beyond the 600 data collections we have decided to preserve to see what else is out there. It is a study of major Federal databases being conducted by the National Academy of Public Administration (NAPA). This study has some interesting exclusions. First of all, we told NAPA not to bother with systems used for generic housekeeping functions, such as personnel, payroll, procurement and supply, because there is little likelihood that we would have any interest in preserving data from such a system. Secondly, we told them not to look at big science, because that is such a large and complex area that it deserves separate attention. (We hope to engage in a project with the National Academy of Sciences on the preservation of scientific data.) Thirdly, we told NAPA not to worry too much about databases on PCs, simply because they would never finish the project if they tried to find all the interesting databases sitting on desktops. With those limitations, NAPA has identified over 10,000 databases.

Obviously, that is far too big a number even for us to think about. So we gave NAPA a set of criteria for culling from the total inventory a subset of those databases with some likelihood that the National Archives would be interested in preserving them. We thought we might wind up with a list of the 500 most important databases in the Federal Government, from an archival perspective. That list would pose quite a challenge for us, because it could practically double the total number of data collections generating data that we want to preserve. The subset of 500 currently has about 900 members.

The next phase of this study is to solicit advice from subject area experts about what data we should try to preserve. NAPA has organized five working groups, with a total of 32 experts in a variety of fields. We are bringing these people together at the end of July for a four day meeting where they will try to develop some common opinions on the long term value of the data.

Which brings me back to the basic point here: what we are dealing with is incoherent data. It concerns practically any area in which the United States Government is involved, which is practically anything. The data we already have ranges from data about tektites on the ocean floor to military operations in time of war. In includes census data on population and the economy, data on Japanese-American internees in World War II, detailed data on air traffic and on stock and bond transactions, and on many, many other subjects. The variety of subjects covered is also increasing.

The data is extremely diverse in content, but content is often the only thing we know about the data until it comes in. We know how many transfers are due, but most often we do not know what the volume of data in a transfer will be, or how it will be organized, even at the physical file level. For example, the files which came in during the first six months of this fiscal year ranged in size from 6 K to 1.4 gigabytes. The number of files in a transfer has ranged from one to 400, and we expect some transfers in the next few years will contain thousands of files.

One thing we do know about the data before it arrives is its logical structure: everything we receive is in flat file format, because we require it to come in in that form. However, we realize that this requirement is unreasonable and unrealistic in many cases. We are working to expand the range of formats we will accept to include relational tables. We expect to change

our regulation to that effect by the end of this year. We know that, when we do that, it will be only one of many steps we will have to take in a journey with no foreseeable end.

That is a brief overview of one aspect of the unique situation of the National Archives. The second aspect is that we are charged with satisfying unknown and unknowable requirements.

NARA's mission to preserve and provide access to records with enduring value makes NARA, in effect, the agent of generations yet unborn. What differentiates this agency from other parts of the government is the unique responsibility NARA has to serve the information needs of the distant future. This responsibility is fundamental to the very essence of the National Archives as keeper of the Nation's memory.

NARA's responsibility to the future places us in a perpetual quandary: we must devote ourselves to serving needs which we cannot know. We cannot know the questions the future will ask of its past, nor how future researchers will go about answering these questions. We must assume, however, that the information technology which will be available in the future --- even in the very near future --- will be more powerful and more flexible than what is available today. Information processing problems which today are difficult and costly, if not impossible, to solve will become as simple as getting a computer to print out narrative in paragraph form. (A short 20 years ago that was beyond state of the art.)

Along with the technology, analytic tools will continue to improve: there will be further developments as powerful as the mathematics of chaos which will help researchers to understand things which today appear to defy reason. We can also assume that events will happen in the future, which will be as

threatening as the depletion of atmospheric ozone, or as exciting as Operation Desert Storm, or as commonplace as the passing of generations, which will make future users want to go back to reexamine the records of the past.